大规模非限定领域汉英双语语料库建设及句子对齐研究*

刘非凡 赵军 徐波

中国科学院自动化研究所 模式识别国家重点实验室 北京 100080

ffliu@nlpr.ia.ac.cn, jzhao@nlpr.ia.ac.cn, xubo@hitic.ia.ac.cn

摘 要： 双语语料库建设及其自动对齐研究对计算语言学的发展具有重要的意义。目前国内外已有的双语语料库尤其是汉英双语语料库规模不大，加工规范不统一，没有形成能够公开使用的通用双语语料库。本文工作在国家 973 子课题支持下，遵循中文语言资源联盟（ChineseLDC）资源共享的宗旨，参照都柏林核元数据元素集制定了双语语料文本标注规范，并对非限定领域双语句子自动对齐技术进行了研究，为大规模建立具有统一标准和规范的、多领域、多体裁、句子级对齐的双语语言信息和知识库奠定了坚实的基础。

关键词： 双语语料库，对齐，中文语言资源联盟

Building Large-Scale Domain Independent Chinese-English Bilingual Corpus and the Researches on Sentence Alignment

Liu Feifan Zhao Jun Xu Bo

National Laboratory of Pattern Recognition

Institute of Automation, Chinese Academy of Sciences, Beijing, 100080

ffliu@nlpr.ia.ac.cn, jzhao@nlpr.ia.ac.cn, xubo@hitic.ia.ac.cn

Abstract: Bilingual corpus shows very important significance in the field of computational linguistics. Currently, the bilingual corpora, especially for the Chinese-English bilingual corpora, are small in scale and nonstandard in the annotation format. Therefore, they can't serve as an open and generally used corpora. Under the support of the-973-project, the paper introduces the specifications for annotating the bilingual text based on Dublin Core Element Set. The paper also presents our researches on the sentence alignment technology for domain-independent bilingual corpus.

Keywords: bilingual corpus, alignment, Chinese Linguistic Data Consortium.

1 引言

国内外很多研究机构都致力于双语语料库的建设，并利用这些语料库进行广泛的研究。加拿大的议会会议录（Canadian Hansards）是非常著名的英法双语语料库，许多最初的基于双语语料库的研究都是在该语料库基础上进行的[1][2]。有关汉外双语语料库建设及其研究，香港科技大学收集和加工了香港立法委员会的会议记录，形成汉英双语语料库[3]。此外，北

*本文受国家 973 项目子课题（G1998030501A-06）和国家自然科学基金项目（60272041）资助。

京大学、东北大学、哈尔滨工业大学的研究人员也建立了一定规模的汉英双语语料库[7][8][9]。但目前汉外双语语料库规模比较小，加工规范也不统一，从而影响了双语语料库知识获取的研究。

实现各个层次的对齐是双语语料库建设的一项重要内容。本文主要讨论汉英双语句子级对齐技术。句子对齐方法基本可以分为三类：

◆ 基于长度的方法：最初由 Brown[1]和 Gale[2]提出，其依据是两种语言译文的长度满足一定比例关系。他们在英法双语的加拿大议会会议录上取得了较好的对齐效果；清华大学和哈尔滨工业大学的研究人员分别将基于长度的方法应用于 Microsoft NT 3.5 Server 安装指南和法律文献的汉英双语句子对齐，获得的试验结果。

◆ 基于词汇的方法：Kay[4]和 Chen[5]则分别根据双语单词的分布信息和词汇翻译模型进行了英德和英法双语句子对齐。文献[8]直接利用双语词典对大学英语教材做了句子对齐，也取得了令人满意的效果。

◆ 混合方法：基于长度的对齐方法模型简单，独立于语言知识和其他外部资源，但鲁棒性不好，容易造成错误蔓延。基于词汇的对齐方法相对可靠精确，但计算相当复杂。研究人员试图将这两种方法结合起来进行句子对齐。香港大学 Wu[3]通过创建特殊词表来对基于长度方法进行了改进，并对在香港立法委员会会议记录上做了对齐试验，取得较好结果。

以上对齐研究大都是围绕单一领域或者某一文献、手册的双语文本进行，本文工作面向多领域多体裁，采用基于双语词典的句子对齐方法进行了文本对齐，并对如何提高对齐精度做了进一步的研究和探讨。该方法不同于 Kay[4]和 Chen[5]的利用译词分布相关性和词汇翻译模型的方法，与文献[8]的对齐方法在评价函数设计、双语词典资源整理上也存在不同之处。

2 双语语料的收集和预处理

原始双语语料主要源自因特网，题材涉及新闻、体育、政治、生活、法律、环境、教育等多个领域，体裁包括演讲、政府报告、报纸、小说、白皮书、答记者问等多种形式。原始语料含有大量的冗余信息和文本噪声，我们对原始语料首先进行了人工预处理，主要包括固定编排格式、统一存储格式、繁体转简体、消除冗余信息和噪声、段落对齐等工作，目前经过预处理后双语语料规模（纯文本格式）约 15M 字节，8 万句左右。

3 双语句子级对齐文本的标注规范

对于中文信息处理各个层面上所需要的语言资源，必须有一套统一标准和规范才能实现资源共享。为更好的与国际接轨，我们参照都柏林核元数据元素集，制定了《双语语料库标注规范（初稿）》，并在 973 标准讨论会上通过专家讨论、审核。规范的标注格式选用 XML 语言，包括以下两部分。

3.1.1 文件头信息

文件头信息就是该双语文本的整体属性信息，参照 Dublin Core Element Set 我们定义了 14 个数据单元，其标记形式和含义见下表。

表一：文件头元素集标记形式及含义

元素	标记形式（XML）	含义
文件头	<TEXT_HEAD>…</TEXT_HEAD>	标明双语文本的文件头
资源名	<TITLE>xx yy#zz</TITLE >	xx 为文件的唯一标识号；yy、zz 表示文本的两种语言标题； # 为分隔符。
作者	<AUTHOR> xx</AUTHOR >	xx 为原文的作者名字
主题词和关键字	<SUBJECT> xx</SUBJECT >	xx 为对语料内容的主题描述
说明	<DESCRIPTION>xx</DESCRIPTION>	xx 是对语料资源内容的描述
资源类型	<TYPE>02020</TYPE>	标识语料资源内容的特征和类型，第一位为创建者编号；第二位表示对齐粒度；第三、四位表示两种语言；第五位表示语体。
来源	<SOURCE>xx</SOURCE>	语料来源
创建者	<CREATOR>xx</CREATOR>	语料加工的主要责任者
出版者	<PUBLISHER>xx</PUBLISHER>	使语料资源成为可获得状态的责任者名称
其他责任者	<CONTRIBUTOR>xx</CONTRIBUTOR>	xx 是除创建者以外，对语料加工做出贡献（包括搜集、翻译、校对等）的其他责任者。
权限	<RIGHTS>xx</RIGHTS>	对语料资源的权限管理声明。
日期	<DATE>yyyy-mm-dd</DATE>	语料资源的创建日期。
格式	<FORMAT>xx</FORMAT>	语料资源的媒体类型。
资源标识符	<IDENTIFIER>xx</IDENTIFIER>	xx 是为获得该资源提供的一个标识符
地区	<AREA>xx</AREA>	原始语料地域

3.1.2 文件体信息

文件体信息包括双语篇章级结构、段落、句子对齐信息。各标记及其含义见表二。

表二：文件体各项标记形式及含义

标记	XML 表示	含义
文件体	<TEXT_BODY>…</TEXT_BODY>	中英文双语篇章
中文标记	<cn>…</cn>	中文篇章
英文标记	<en>…</en>	英文篇章
段落标记	…	xx 为文本中段落标号
句对标记	…	xx 为句对在文本中的标号；yy 为句对中含有的句子数。
句子标记	<s id=xx>…</s>	xx 为句子在本段落中的标号。

4 基于双语词典的中英双语句子对齐

本文采用了基于双语词典的对齐方法，通过设计合理的评价函数计算双语句子之间互为译文的评价值，最后运用动态规划算法搜索整体评价值最高的句对序列。

4.1 句子对齐的形式化描述

本文的句子对齐是在段落对齐的基础上进行的。用 $Para<C, E>$表示一个包含 m 句中文和 n 句英文的双语段落，其中 C 为 m 个中文句子（$sc_1 \cdots sc_i \cdots sc_m$）组成的句子集合，$E$ 为 n 个英文句子($se_1 \cdots se_j \cdots se_n$)组成的句子集合，$sc_i$ 表示第 i 个中文句子，se_j 表示第 j 个英文句子。那么 $\forall E_a \subset E, C_a \subset C$ （E_a 和 C_a 不可同时为空集），$a=<C_a, E_a>$构成一个双语句对。根据含有的中文句子个数（$|C_a|$）和英文句子个数（$|E_a|$），双语句对可以分为空对一、一对空、一对一、多对一、一对多、多对多等六种类型。一个双语段落内部存在许多种双语句对组合，每一种双语句对组合代表一种对齐方式。句子对齐就是要在所有的双语句对组合中搜索一个最佳双语句对序列，即获得一个最佳对齐方式 $A=a_1 \cdots a_i \cdots a_r$（其中 $a_i=<C_{ai}, E_{ar}>$表示第 i 个双语句对，r 为双语句对个数），该最佳对齐方式中各句对要满足以下条件：

- ✧ 完备正交性：$\begin{cases} \bigcup_{i=1}^{r} E_{ai} = E \quad \bigcup_{i=1}^{r} C_{ai} = C \\ \forall 1 \le i, j \le r, i \ne j \quad E_{ai} \cap E_{aj} = \phi \quad C_{ai} \cap C_{aj} = \phi \end{cases}$
- ✧ 无交叉性：$\forall 1 \le i < j \le r$ $\begin{cases} 若 se_u \in E_{ai}, se_v \in E_{aj} \quad 必有 u < v \\ 若 sc_u \in C_{ai}, sc_v \in C_{aj} \quad 必有 u < v \end{cases}$
- ✧ 互译匹配最优性：该双语句对序列整体的互译匹配度优于满足以上条件的其他双语句对序列。
- ✧ 不可分割性：任何一个 $\forall a_i \in A\ (1 \le i \le r)$ 都不能再分解成两个或者多个更小的符合上述条件的句对。

句对序列的互译匹配程度用一个评价函数 S 来衡量，每个可能的句对序列都有一个评价值 $S(A_i)$，那么句子对齐问题即转化为下列最优化问题：

$$A = \arg\max_{1 \le i \le k} S(A_i) \quad (k \text{ 为可能出现的句对序列数}) \qquad (1)$$

4.2 句对序列评价函数和双语句对内部互译匹配评价函数

设计一个恰当的评价函数来衡量一个句对序列整体互译对应程度是基于双语词典句子对齐方法的核心问题。本文中一个句对序列的评价值由该序列中每个句对的评价函数值的代数和来获得。假设第 i 个句对序列有 h 个句对 A_i（$a_1 \cdots a_h$），则该句对序列的评价值为：

$$S(A_i) = \sum_{j=1}^{h} Score\ (a_j) \qquad (2)$$

式中 $Score$（a_j）为双语句对 a_j（$1 \le j \le h$）的评价函数，用来评价句对内部的互译匹配度。

评价句对内部的互译匹配程度就是考察该句对所含中英文句子之间的词语匹配信息，文献[8]采用从英文单词向中文句子匹配的方法，虽然避免了分词带来的错误，但是由于汉语的特殊性和复杂性、语料的领域广泛性，很容易造成误匹配。比如“certainly”译文为“的确”，在句子“他的确切地址……”中便可以很好的匹配。

本文从中文到英文的匹配的角度来研究句对内部对齐的衡量尺度。考察一个双语句对 $a=<C_a,E_a>$，用 x、y 分别表示该句对中含有的中文句子和英文句子个数，p、q 分别表示句对中含有的中、英文词数。匹配具体算法如下：

1）对每个中文单词 c_i（$0\le i\le p$），利用双语词典查找相应的英文翻译列表 $T_i(t_1,t_2,\cdots t_f)$，f 为双语词典中 c_i 的英文翻译词条数目；

2）对英文所有的单词 e_j（$0\le j\le q$）进行词形还原；

3）利用下式计算中文单词 c_i 和 e_j 的匹配评价：

$$Match(c_i,e_j)=\underset{1\le k\le f}{Max}\left(\frac{Max\,\Pr eMatch(t_k,e_j)}{Max(Len(t_k),Len(e_j))}\right) \tag{3}$$

其中 $Max\,\Pr eMatch(t_k,e_j)$ 求取 t_k 和 e_j 从左侧起最大匹配的字母数，$Len(e_j)$ 为词汇 e_j 的长度。如果 t_k 和 e_j 完全一致，显然将返回 1，否则返回一个 0 到 1 之间的小数。该公式会引入一些冗余对应信息，本文采用了 0.7 作为阈值，小于这个阈值则舍弃。

4）双语句对 a 的整体评价值可以由下式求得：

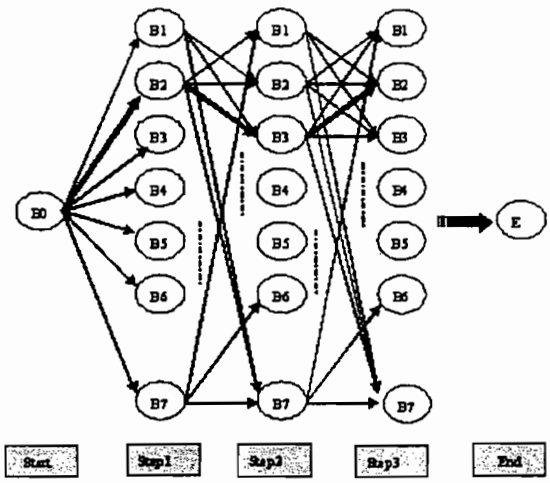
$$Score(a)=\frac{\sum_{i=1}^{p}\underset{1\le j\le q}{Max}(Match(c_i,e_j))}{\Psi} \tag{4}$$

其中 Ψ 是一个归一化因子，可以选取 a：$p+q$　b：p^2+q^2　c：$\sqrt{p^2+q^2}$。

5）由公式（2）计算该句对序列评价值，并通过式(1)实现对齐。

4.3　最优句对序列的动态规划搜索

本文考虑了 7 种对齐类型：（$B1$）1－0；（$B2$）0－1；（$B3$）1－1；（$B4$）1－2；（$B5$）2－1；（$B6$）1－3；（$B7$）3－1。图一中每个阶段有 7 个节点 $B_i(1\le i\le 7)$ 分别代表 7 种句对类型，搜索最优句对序列就是寻找一条最优路径，使得沿着这条路径获得的句对序列的评价值最高（图中粗线显示），可以按照下面的递归公式来实现：



图一 最优句对序列动态规划搜索

$$\begin{cases} M_1(B_j)=Measure_{\,0}(B_0,B_j) & s=1 \\ M_s(B_j)=\underset{i}{Max}(M_{s-1}(B_i)+Measure_{\,s-1}(B_i,B_j)) & s>1 \end{cases} \tag{5}$$

$$Path_s(B_j)=\arg\max_i(M_{s-1}(B_i)+Measure_{\,s-1}(B_i,B_j)) \quad s>1 \tag{6}$$

其中 $M_s(B_j)$ $(1\le j\le 7)$为第 s 步到达节点 B_j 遍历的所有路径中句对评价值总和的最大

值，$Measure_{s-1}(B_i, B_j)$ 表示由 $s-1$ 步的节点 B_i 到达节点 B_j 需要考察的句对的评价值，可以利用前面讨论的句对评价函数 $Score(a)$ 来计算。公式（6）用来保存每一步的回溯路径。

4.4 句对内部匹配评价函数的改进

上面讨论的匹配算法中，有限的双语词典覆盖度影响了句对内部匹配度的计算，本文尝试引入汉语《同义词词林》和英文 Wordnet 来扩展双语词典的覆盖度，并加入了结构匹配和对中文数字的专门处理。

➢ 增加结构匹配信息

汉语中一些常用的形容词、副词，经分词后“的”、“地”都作为单独一个词，这样匹配中会导致这些词在双语词典中查询英文翻译失败。改进算法在匹配中文词的时，如果后一词为“的”、“地”，而且两个词合起来在双语词典中作为一个词出现，则合为一个词进行匹配。

在汉语词的英文翻译词条是一个词组的情况下，4.2 的匹配算法不能有效处理。在改进算法中，英文匹配窗口由一个词扩展到三个词，从而在一定程度上可以解决词组匹配的问题。

➢ 引入《同义词词林》和 Wordnet 扩展双语词典的覆盖度

不同领域的语料对同一个意思的表达方式都不一致，双语词典的覆盖度很难满足要求。引入《同义词词林》和 Wordnet 进行同义词扩展在一定程度上可以弥补上述不足。

➢ 增加对全角数字的专门处理

中文数字在很多语料中是全角形式出现的，导致匹配失败。对其进行专门处理就是在匹配前先将数字转换为半角字符，直接和英文进行数字匹配，如果匹配不上利用 Wordnet 扩展后再进行匹配。

5 试验结果和分析

试验语料是从收集整理的双语语料中选取的 9 个文本（表三），实际含有 1996 个句对。

表三：试验语料的领域体裁分布表

	文本 1	文本 2	文本 3	文本 4	文本 5	文本 6	文本 7	文本 8	文本 9
领域	文化	政治	文化	政治	经济	政治	经济	政治	法律
体裁	报纸新闻	报纸新闻	报纸新闻	政府报告	演讲	政府报告	演讲	答记者问	法律条文

由于系统性能与召回率和精确度均有关，本文采用 F一测试作为最后的评价结果。

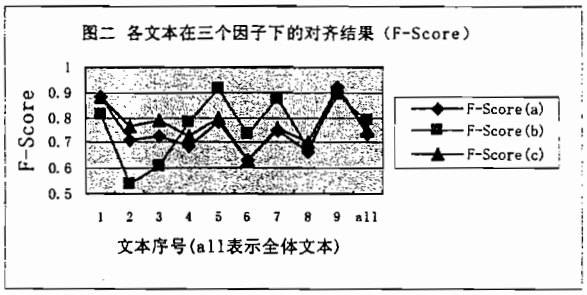
$$precision = \frac{\text{正确识别的句对数}}{\text{识别出来的句对总数}} \qquad recall = \frac{\text{正确识别的句对数}}{\text{实际含有的句对总数}} \qquad (7)$$

$$F = \frac{(\beta^2 + 1) \times precision \times recall}{\beta^2 \times precision + recall} \qquad \text{（本文试验中 } \beta = 1\text{）} \qquad (8)$$

5.1 试验一：归一化因子的选取

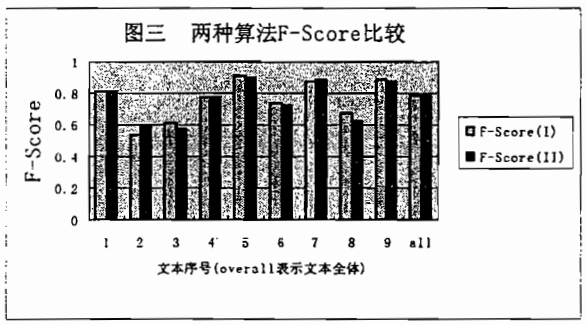
在原算法（算法一）的基础上，本文对公式（4）中三个归一化因子选取进行了试验分析。由图二可以看出，对于不同领域和体裁的语料，三个因子反应趋势相似，但敏感度不同，这也说明多领域双语语料句子对齐和限定领域相比复杂很多。实际上，归一化因子也起到了

惩罚因子的作用，因子 b 的惩罚能力较强，更能有效的抑制一对多、多对一的情况，所以在整体上取得了相对较高的精确度和召回率，而因子 a 和 c 仅在个别文本中可以达到比较高的精确度和召回率。因此本文选用归一化因子 b。



5.2 试验二：两种算法的比较

运用改进算法（算法二）重新对试验语料进行对齐，从图三可以看出，改进的算法二并不十分理想，仅对个别领域文本的对齐结果有些改善。主要原因可能有：



- 语料的复杂和不规范性；
- 繁转简、分词等预处理过程中引入的一些系统错误或人为错误无法通过双语词典的扩展来校正。
- 一词多义现象使同义词扩展引入一些额外的冗余信息，一定程度上影响了对齐精度。
- 汉语《同义词词林》电子词典在收录上也存在一些问题，对其进行进一步加工和整理，可能会获得比较理想的效果。

6 结束语

本文在制定双语对齐文本标注规范的基础上，收集整理了大量的中英双语语料，运用基于双语词典的方法实现了句子对齐，并对提高对齐精度做了进一步的研究。本文提出的对齐方法与人工校对相结合，对于实现非限定领域双语语料的句子对齐是比较适合和实用的。

参 考 文 献

[1] P.F.Brown, J. C. Lai & R. L. Mercer: Aligning Sentences in Parallel Corpora, *ACL-29*, 169-176, 1991

[2] Gale, Church: A Program for Aligning Sentences in Bilingual Corpora, *Computational Linguistics 19(1)*, 1991

[3] Dekai Wu: Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria, In *ACL-94*: 80-87

[4] M. Kay & K. Roescheisen: Text-Translation Alignment, *Computational Linguistics 19(1)*,121-142, 1993

[5] S. F. Chen: Aligning Sentences in Bilingual Corpora Using Lexical Information, *ACL-31*, 1993

[6] 刘昕，周明，黄昌宁．基于长度算法的中－英双语文本对齐的试验．计算语言学进展与应用．1995

[7] 吕学强，李清隐，陈文亮，姚天顺．汉英法律文献的子条级自动索引和对齐．中文信息学报 2002（4）

[8] 杨沐昀，李生，赵铁军，方高林，吕雅娟．A Research on Bilingual Dictionary Based Sentence Alignment for Chinese English Parallel Corpus. 高技术通讯（英文版）．2002，8（2）

[9] 常宝宝 詹卫东 柏晓静 吴云芳 张化瑞．服务于汉英机器翻译的双语语料库和短语库建设．第二届中日自然语言处理专家研讨会论文集．2002．p147-154.